DECISION THEORY WITHOUT LUMINOSITY

YOAAV ISAACS AND BEN LEVINSTEIN

Abstract. Our decision-theoretic states are not luminous. We are imperfectly reliable at identifying our own credences, utilities, and available acts, and thus can never be more than imperfectly reliable at identifying the prescriptions of decision theory. The lack of luminosity affords decision theory a remarkable opportunity—to issue guidance on the basis of epistemically inaccessible facts. We show how a decision theory can guarantee action in accordance with contingent truths about which an agent is arbitrarily uncertain. It may seem that such advantages would require dubiously adverting to externalist facts that go beyond the internalism of traditional decision theory, but this is not so. Using only the standard repertoire of decision-theoretic tools, we show how to modify existing decision theories to take advantage of this opportunity.

These improved decision theories require agents to maximize conditional expected utility—expected utility conditional upon an agent's actual decision-situation. We call such modified decision theories "self-confident". These self-confident decision theories have a distinct advantage over standard decision theories—their prescriptions are better.

1. Introduction

Our decision-theoretic states are not luminous. We are imperfectly reliable at identifying our own credences, utilities, and available acts, and thus can never be more than imperfectly reliable at identifying the prescriptions of decision theory. Even if we figured out what the one true decision theory is, we could follow it only unreliably.

The lack of luminosity is cause for dejection, as there are some ways in which decision theory cannot work out as well as one might have expected. But the lack of luminosity is also cause for elation, as there are some ways in which decision theory can work out better than one might have expected. The lack of luminosity affords decision theory a remarkable opportunity—to issue guidance on the basis of epistemically inaccessible facts. We show how a decision theory can guarantee action in accordance with contingent truths about which an agent is arbitrarily uncertain. It may seem that such advantages would require dubiously adverting to externalist facts that go beyond the internalism of traditional decision theory, but this is not so. Using only the standard repertoire of decision-theoretic tools, we show how to modify existing decision theories to take advantage of this opportunity. Rather than arguing for causal decision theory against evidential decision theory or vice versa, we improve both causal decision theory and evidential decision theory.

These improved decision theories require agents to maximize *conditional* expected utility—expected utility conditional upon an agent's actual decision-situation. We call such modified decision theories "self-confident". These self-confident decision theories have a distinct advantage over standard decision theories—their prescriptions are better.

This advantage comes at a price. Self-confident decision theories are sometimes hard to follow. Their key prescriptions may well seem alien. But all decision theories are sometimes hard to follow. All decision theories sometimes issue alien prescriptions. Without luminosity, that's just how decision theory has to go. The price paid for self-confident decision theories is a price that must be paid anyway—we may as well get some payoff for it. The distinctive feature of self-confident decision theories is not that their prescriptions are non-luminous, but is rather that they capitalize on the non-luminosity of their prescriptions.

We show that self-confident decision theories are not strictly harder to follow than standard decision theories. While there are some circumstances in which standard decision theories are easier to follow, there are also some circumstances in which self-confident decision theories are easier to follow.

In light of these advantages, we maintain that causal decision theory should be superseded by self-confident causal decision theory and that evidential decision theory should be superseded by self-confident evidential decision theory. The correct decision theory—whatever else it is—is self-confident.

2. Using Decision Theory

Decision theory specifies which of an agent's available acts are rational given that agent's credences and utilities. For this reason, understanding decision theory can be very useful. If an agent is certain about what her credences, utilities, and available acts are and certain which available acts are rational given any credences, utilities, and available acts, then probabilistic coherence guarantees certainty about which of her available acts are rational.

But let's not get too excited about how useful understanding of decision theory is. This guaranteed certainty about which available acts are rational required certainty of credences, utilities, and available acts. And an agent can be uncertain about those. Given such uncertainty, even an unimpeachable understanding of decision theory may leave an agent uncertain as to what she ought to do.

Uncertainty about an agent's own credences, utilities, and available acts is not only possible, it is actual. Perfect self-knowledge is a lovely thing—too lovely to be easy, let alone guaranteed. It would be outlandish to claim that all agents are always completely certain about their credences, utilities, and available acts. There is insight in E.M. Forster's line, "How can I tell what I think till I see what I say?"[1] We can't always tell what we think, what we want, and what we can do. In a sudden crisis, and agent may be unable to determine what she thinks or what she wants.[2] Moreover, recent work on implicit bias has revealed just how much ignorance we can have about our own beliefs.[3] And the case that we understand ourselves imperfectly becomes even more forceful when we think in decision-theoretic terms, about credences and utilities. Credences can have any real value between 0 and 1, and no one can introspect well enough to reliably feel the difference between credence .523924357 and credence .523924358.[4] Only an arrestingly extreme Cartesianism would claim that all agents are always completely certain about their credences, utilities, and available acts. Such views are increasingly disfavored.

Following Williamson (2000), we will say that a condition which is such that an agent can know that it obtains whenever it obtains is *luminous.* Williamson argues that there are no non-trivial luminous conditions.[5] Generalizing his argumentation to probabilistic contexts, Williamson argues that no non-trivial condition is such that it has evidential probability 1 whenever it obtains.[6] He dubs a condition that can always have probability 1 *luminous in probability 1.* It would be easier to follow the dictates of decision theory if credences, utilities, and available acts were luminous and luminous in probability 1.[7] But they're not. They're not even close.

It would be a real problem if decision theory fell silent when an agent was uncertain about her credences, utilities, and available acts, for then decision theory would fall silent regarding literally every decision every ordinary person ever made. But there is no such reason for decision theory to fall silent. There is nothing formally problematic about the lack of luminosity. Extant decision theories require

[1] Forster (1985).

[2] See Kagan (2018), and Lasonen-Aarnio (ming). See also Spencer and Wells (2019).

[3] See Gendler (2011) for more.

[4] See Carr (2020). A similar issue applies to utilities, but stating that issue is somewhat more complex since utility values are defined only up to positive affine transformation.

[5] For previous argumentation against Cartesianism, see Bonjour (1980) and Peacocke (1998).

[6] See Williamson (2008).

[7] The arguments against a condition being luminous and against it being luminous in probability 1 are nearly identical. For simplicity, we will use the term "luminous" broadly, as meaning "both luminous and luminous in probability 1".

credences, utilities, and available acts, but do not require certainty about them. In fact, some extant decision theories make particular use out of intermediate credences about credences, utilities, and available acts.[8] The lack of luminosity doesn't keep decision theory from issuing norms; it just makes those norms harder to follow. That's not so bad. But there's a bigger problem. The dictates of standard decision theories are sometimes wrong.

3. Against Futility

A rational decision can go badly. Sometimes a good plan is frustrated. Sometimes a good bet is lost. The world is sometimes uncongenial—so it goes.

A rational decision can go badly, but a rational decision cannot have to go badly. There must be some way for a good plan to succeed. There must be some way for a good bet to be won. The world is sometimes uncongenial, but it can't have to be uncongenial. A rational agent cannot face a decision situation—a triple of a credence function, utility function, and a set of available acts—in which her decision has to go badly. If a decision has to go badly—if an agent's credences, utilities, and available acts entail that a choice will go badly—then it is a bad choice. We contend that the following principle is true:[9]

> **Anti-Futility:** In any choice between A and B, a rational agent will not choose A if—holding the facts of that decision-situation fixed—choosing A must lead to a worse outcome than choosing B.

Note that Anti-Futility is not a trivial consequence of standard decision theories. It might seem like a mere restatement of dominance reasoning, but it isn't. Dominance reasoning concerns superiority across all states, whereas Anti-Futility only concerns superiority given the agent's decision situation. Anti-Futility is thus a stronger principle, and moreover is inconsistent with standard decision theories. Standard decision theories prescribe that agents maximize expected utility.

> **Expected Utility Maximization:** In any choice between A and B, a rational agent will choose A if choosing A has greater expected utility than choosing B.[10]

If agents always act so as to maximize their expected utilities, they will sometimes violate Anti-Futility. Most starkly, if agents always act so as to maximize their expected utilities, they will sometimes take bets that they are guaranteed to lose.

[8] See Joyce (2002) and Hájek (2016) for more.

[9] For convenience, we restrict our attention to decision situations with two available acts.

[10] There are different sorts of expected utility—causal expected utility, evidential expected utility, and so on. But the differences among those expected utilities are immaterial here.

It may seem outlandish that Anti-Futility and Expected Utility Maximization should conflict in this way. After all, one of the most celebrated results in decision theory is that an agent is representable as an expected utility maximizer if and only if she cannot be (synchronically) Dutch Booked.[11] As we'll see below, the traditional formal notion of a Dutch Book isn't general enough for evaluating anti-luminous decision-making. For now, suffice it to say that Dutch Books require a guaranteed loss in all possible worlds. In anti-luminous situations, some possible worlds are ones where the agent's decision situation is different from how it actually is. However, violating Anti-Futility requires only a guaranteed loss over the worlds where the agent's decision situation is as it actually is. In other words, a set of bets violates Anti-Futility if it guarantees a loss given the *actual truth-values* of sentences about the agent's decision-situation and any consistent assignment of truth-values to the other sentences in the agent's language.[12]

An agent who maximizes expected utility will violate Anti-Futility only when that agent is uncertain about her decision-situation. So why does an agent's ignorance about her decision-situation cause problems which his ignorance about other matters doesn't? Some cases will clarify the matter:

> **Case 1:** Suppose that Jane has credence $\frac{1}{2}$ that it will rain. Jane is then offered a bet about the rain at favorable odds—she'll gain \$2 if it doesn't rain and only lose \$1 if it rains. Jane takes the bet. But it rains, and so Jane loses \$1.

Has Jane violated Anti-Futility? Not at all. Jane lost the bet, but she was not guaranteed to lose the bet. The very same decision-situation could have occurred in a world in which it didn't rain, and then Jane would have won the bet.[13]

> **Case 2:** Suppose that Jane has credence $\frac{1}{2}$ that it will rain. But Jane does not know her own mind, at least not perfectly. Jane has credence $\frac{1}{2}$ that she has credence $\frac{1}{2}$ that it will rain. Jane is then offered a bet about her credences at favorable odds—she'll gain \$2 if she doesn't have credence $\frac{1}{2}$ that it will rain and only lose \$1 if she has credence $\frac{1}{2}$ that it will rain. Jane takes the bet. But Jane has credence $\frac{1}{2}$ that it will rain, and so Jane loses \$1.

Has Jane violated Anti-Futility? Yes. Not only did Jane lose the bet, she was guaranteed to lose the bet. The very same decision-situation could have occurred in various other worlds, but in each of those worlds Jane has credence $\frac{1}{2}$ that it will rain, and in each of those worlds Jane loses the bet.

[11] See Ramsey (1926).

[12] See Mahtani (2015) for more about the formal underpinnings of Dutch books.

[13] That is, there are worlds in which an agent with Jane's credences, utilities, available acts, and decision function wins the bet that it won't rain.

The problem is this: the outside world can vary freely from Jane's decision-situation, but Jane's decision-situation cannot vary freely from itself. Jane's behavior in the first decision-situation did not guarantee that she would lose the bet about the rain. But Jane's behavior in the second decision-situation did guarantee that she would lose the bet about her decision-situation.

As case 2 shows, Anti-Futility conflicts with Expected Utility Maximization. So which principle should we hold to? Expected Utility Maximization does have a fine pedigree. Expected Utility Maximization—unlike Anti-Futility—is a foundational principle of decision theory. But Anti-Futility is a very appealing principle. It is upsettingly strange to think that the decision-theoretic underpinnings of a rational bet could guarantee the loss of that bet. While there are times when one must learn to live with such upsetting strangeness, there are also times when one must learn to avoid it. There is enough to be said for Anti-Futility that it may be worth revising decision theory to satisfy it.

4. Self-Confidence

We've seen that Anti-Futility and Expected Utility Maximization are inconsistent. An agent who maximizes expected utility and is uncertain about her credences, utilities, or available acts will take some bets that—as she is—she cannot win. Given expected utility maximization, any uncertainty about an agent's decision-situation can be exploited by a bet about the agent's decision-situation. After all, since the decision-situation is the same in every world in the decision-situation, any uncertainty about the decision-situation is exploitable in the same way as uncertainty about a tautology.

We have two choices. We can give up on Expected Utility Maximization, or we can give up on Anti-Futility. Suppose we give up on the former. What do we do instead? What sort of decision theory can satisfy Anti-Futility? It's easiest to think about a specific case first. Let's imagine a decision theory which (somehow) obeys Anti-Futility. We'll call it "T-theory". Think back to Jane, who was uncertain about her credences about the rain. How can we make sense of a T-theory which would forbid Jane from taking a losing bet about her credences, even at apparently favorable odds?

Like all decision theories, T-theory makes its prescriptions in light of Jane's credences. But in light of Jane's credences the bet is a bad idea. After all, Jane's credences guarantee that the bet will lose. Jane's credence about her credence about rain doesn't entail that the bet is a bad idea, but Jane's credence about the rain does entail that the bet is a bad idea. And nothing prevents T-theory from taking Jane's credence about the rain into account. There's valuable information available for T-theory to capitalize on that standard decision theories unwisely ignore.

Any decision theory which obeys Anti-Futility must take an agent's credences, utilities, and available acts for granted and ignore any possibilities in which they are other than as they actually are. In order to obey Anti-Futility, a decision theory must act as though the only possibilities of consequence are those in which the agent's decision-situation is as it actually is. And this can be done—a decision theory has all the information required to ignore possibilities in which the decision-situation is other than it actually is. After all, a decision theory specifies a way of taking an agent's credences, utilities, and available acts, and uses those to pick an action. A decision theory can respond to more information about a decision-situation than an agent is aware of.

We can formulate decision theories which ignore all possibilities in which the agent's decision-situation is other than it actually is. We will call such decision theories "self-confident". What self-confident decision theories do with possibilities they consider will vary from theory to theory, just as what standard decision theories do with a full space of possibilities varies from theory to theory. There is a self-confident decision theory that treats its possibilities the way causal decision theory treats its possibilities, a self-confident decision theory that treats its possibilities the way evidential decision theory treats its possibilities, and so on. Instead of prescribing acts which maximize some sort of expected utility across all worlds, self-confident decision theories will prescribe acts which maximize some sort of expected utility across worlds in which the agent's decision-situation is what it actually is. Instead of evaluating acts relative to the agent's unconditional credences, self-confident decision theories evaluate acts relative to the agent's credences conditional upon the agent's actual decision-situation.[14]

Regarding matters that are independent of the agent's decision-situation, the self-confident analogue of a standard decision theory will give the same prescriptions as the standard decision theory.[15] For propositions independent of the agent's decision-situation, self-confident decision theories prescribe that an agent with credence of x in p be willing to pay up to x dollars for a bet that pays 1 dollar if p is true and nothing otherwise.[16] But when it comes to matters that are

[14] For an analogous approach to updating, see Schoenfield (2017) contrasting conditionalization with conditionalization*.

[15] The precise character of the independence (evidential, causal, etc.) will depend on which standard decision theory is at stake.

[16] Taking for granted that the agent's utilities are linear in money.

not probabilistically independent of the agent's decision-situation, self-confident decision theories can produce different behavior. Consider the limiting case of dependence: a self-confident decision theory's prescriptions for bets regarding the agent's decision-situation. If the proposition being bet on is that the agent's credences are whatever they actually are, that the agent's utilities are whatever they actually are, that the agent's available acts are whatever they actually are, or that the agent's decision function is whatever it actually is, then a self-confident decision theory will prescribe that the agent be willing to pay up to 1 dollar for a bet that that pays 1 dollar if p is true and nothing otherwise. Put a bit loosely, a self-confident decision theory will have the agent act as though he were supremely confident about his decision-situation, however much doubt the agent may entertain. But the affected confidence that a self-confident decision theory prescribes cannot lead an agent astray; self-confident decision theories always get an agent's decision-situation right. Self-confident decision theories act as though some unfalsified worlds were falsified, but those unfalsified worlds are always false.

4.1. **Some Formalities.**

To model anti-luminous decision problems, we introduce some formal apparatus. First, we should model the agent's uncertainty. Let Ω be a set of epistemically possible worlds. In normal decision problems, one's credence function, utilities, and available acts remain constant across Ω. But here, these properties of the agent vary.

So, for each $\omega \in \Omega$, we let

- c_ω be the agent's credence function at ω,
- u_ω be the agent's utility function at ω,
- A_ω be the agent's *available* acts at ω, and finally
- $d(\omega) = \langle c_\omega, u_\omega, A_\omega \rangle$ is the triple of credences, utilities, and available acts at a world.

We then represent her *generalized decision problem* as an ordered pair $\langle \Omega, d \rangle$. For a given generalized decision problem X, we will also write A^X_ω, c^X_ω, and u^X_ω to make clear that we're referring to the available acts/credences/utilities at a world in a given generalized decision problem.

We can define an equivalence class over Ω where $\omega \sim \omega'$ if and only if $d(\omega) = d(\omega')$. We call $d(\omega)$ the decision situation at ω, and use the notation $[D = d(\omega)]$ to refer to the set of worlds where the decision situation is the same as the one at ω.

If the agent finds herself in $\langle \Omega, d \rangle$, an EU-maximizing theory will tell her to choose the act in A_ω that maximizes the expected value of u_ω according to c_ω for whichever world ω the decision is taking place at.[17] The method of calculating expected utility will differ from theory to theory, but the important point is that the calculation is a function of her actual credences, utilities, and available acts.

More abstractly, if T is an EU-maximizing theory, it will rank all available acts at a given world in terms of T's method of calculating expected utility. The expected utility of an act a itself is calculated using the agent's utility function at a world along with some probability function that is itself a function of just the decision problem and a.

In particular, there is a function f_T such that for any decision problem X, $\omega \in \Omega^X$, and $a \in A^X_\omega$:

- $f_T(\omega, a, X) = P$ for some probability function P over Ω.
- the expected utility of a according to T is

$$\mathrm{EU}_T(a, \omega) = \sum_{\omega' \in \Omega} f_T(\omega, a, X)(\omega') u_\omega(a, \omega')$$

.

Call f_T T's credal map. For example, for standard evidential decision theory, the credal map $f_{\mathrm{EDT}}(\omega, a, X) = c_\omega(- \mid a)$. For standard causal decision theory, $f_{\mathrm{CDT}}(\omega, a, X) = c_\omega(a \longrightarrow -)$.

Self-confident theories calculate expected utility based on the agent's credences conditional on her decision situation. To spell this out, for any generalized decision problem X and world $\omega \in \Omega^X$, we defined $c^\star_\omega := c_\omega(- \mid [D = d(\omega)])$. I.e., $c^\star_\omega$ is the agent's credence function at ω conditional on her actual decision-situation at ω. For a generalized decision problem X, we can define $X^\star$ to be the same as X, except that in $X^\star$, at any world ω, the agent's credences are $c^\star$ instead of c. $X^\star$ is just X with the credences replaced by those conditional on the agent's actual situation at each world. An EU-maximizing theory T is *self-confident* if and only if its credal map f_T is such that $f_T(\omega, a, X) = f_T(\omega, a, X^\star)$.

Note that for any standard EU-maximizing theory that isn't self-confident, we can formulate one that is based solely on facts encoded in the agent's decision-situation. Self-confident CDT and EDT are just like their standard versions, except they use $c^\star$ instead of c to calculate expected utility. Furthermore, a self-confident decision-theory doesn't require the agent to have any epistemic access to her starred-credences. Instead, it merely ranks acts based on those credences, which are fully determined by her unconditional credences, available acts, and utilities at each world.

[17] Alternatively, it may maximize the expected value of some mix of utility functions that vary across Ω, but these differences will not matter for the discussion below.

5. Evaluating Self-Confidence

Self-confidence has obvious advantages. In general, one expects self-confident agents to fare better than non-self-confident agents. Self-confident agents will win bets that non-self-confident agents will lose.

Suppose that you are building a faithful robot servitor, Hal. You can program Hal to have credences, utilities, a set of available acts, and a function from his credences, utilities, and available acts to an act. Unfortunately, you just don't have the materials to make sure that Hal is always certain about his credences, utilities, set of available acts, and decision function. Hal will be insuperably uncertain of his precise decision-situations. Yet despite this limitation, you can program Hal with any decision function you want. If you want Hal to make the best decisions possible, you should make sure that Hal's decision function is self-confident. By giving Hal a self-confident decision function, you can provide Hal with the benefits of the self-knowledge which you cannot actually provide him. But if self-confidence is a good thing for Hal's decision function to have, then it stands to reason that self-confidence is a good thing for our decision functions to have as well.

Let's begin by looking at Savage's decision theory—a simple, classic framework best applicable to cases in which acts and states are (evidentially and causally) independent. This independence assumption guarantees that causal decision theory and evidential decision theory will not diverge from one another—they prescribe maximizing expected utility, and in the same way. This classic, expected utility-maximizing decision theory can be contrasted with self-confident decision theory. And self-confident decision theory has some clear advantages.

According to standard EU-maximizing Savage-style decision theory, if ω is the actual world, and c_ω is the agent's actual credence function, she should choose the available act

$$a^* = \arg\max_{a \in A} \sum_{\omega'} c_\omega(\omega')u(a, \omega').$$

According to self-confident decision theory, she should instead choose:

$$a^* = \arg\max_{a \in A} \sum_{\omega'} c_\omega(\omega' \mid D = d(\omega))u(a, \omega').$$

There's a famous result which entails that any agent should prefer the decisions made by self-confident decision theory to those made by classic decision theory. Good's Theorem (1967) is commonly glossed as telling us that learning free information is always valuable in expectation, but in reality, the theorem is about which credences you'd prefer to use for decision making. No learning need take place. What Good's Theorem really says is that if $\mathcal{E}$ is a partition of Ω, then:

$$\sum_{E\in\mathcal{E}} \max_{a\in A} \mathrm{EU}(a \,|\, E)c(E) \geq \max_{a\in A} \mathrm{EU}(a)$$

with strict inequality whenever $\max_{a\in A} \mathrm{EU}(a \,|\, E) \neq \max_{a\in A} \mathrm{EU}(a)$ for some $E \in \mathcal{E}$. (Here, EU refers to Savage-style expected utility.)

What this means is that, from the point of view of your unconditional credences (which are unaware of which $E \in \mathcal{E}$ is true), you expect your credences conditional on the true cell of $\mathcal{E}$ to make a better choice (or a choice that's at least as good when your mind would never change). You never have to actually *learn* which $E \in \mathcal{E}$ is true. You simply prefer your credences conditional on the true cell—whatever it is—to your unconditional credences.

Note that the cells $D = d(\omega)$ for each $\omega \in \Omega$ form a partition. The self-confident agent does not learn which cell she is in, but she does use her credences conditional on the true cell—whichever cell that is—to calculate expected utility. Thus, by Good's Thoerem, the unconditional EU-maximizer expects that she would do better if she followed self-confident decision theory.[18]

What this shows is that your conditional credences are better guides to the world. Of course, sometimes acts and states are not independent, which is why CDT and EDT were invented. Such dependencies can lead expected utility maximizers to make questionable decisions. When acts and states are independent, an expected utility maximizer will not pay to avoid information. But when acts and states are not independent, followers of EDT will sometimes pay to avoid information (and some have argued that followers of CDT will as well).[19] It's not just that the assumptions used to prove Good's theorem don't hold; the result itself—that free information never has negative value—doesn't hold. So, there's no guarantee that an EDT or CDT agent will *always* prefer to use her conditional rather than unconditional credences.

Nonetheless, we maintain that a decision theory should prefer using conditional credences to unconditional credences. If by the lights of EDT or CDT it's better to use the unconditional credences than the conditional credences, that just shows that those lights are irrational. Straightforwardly, you should think your credences

[18] There is an important nuance worth highlighting here. When we specify your actual credence function *de re*, that credence function always expects its conditional version to do better. However, it is not in general true that your credence function, whatever it is, expects that your credence function, whatever it is, conditioned on true information about its own decision situation will do better. See Dorst et al. (2021).

[19] See Maher (1990).

conditional on true information are epistemically superior. Indeed, on any standard measure of accuracy, you must expect that your credences conditional on true information are at least as accurate as your current credences.[20] Using more accurate credences to decide is generally a good thing. Given that the conditional credences are epistemically superior to the unconditional credences, they cannot credibly be pragmatically inferior.[21]

Furthermore, following an unconditional EU-maximizing theory in anti-luminous situations is often exploitable even when the agent's underlying credences are coherent. Let's return to Jane and bets about the rain. We saw above that, as an EU-maximizer with non-luminous credences, Jane would take bets that she was guaranteed to lose in her decision-situation. But the situation is even worse than that. There is a collection of bets Jane can be offered such that—given the bets she will take as an EU-maximizer—she will lose money in any possible decision situation.

To illustrate, suppose that Jane's credence is either .6 that it will rain or .4 that it will rain. She's 50% sure that her credence is .6 and 50% sure it's .4. There are thus four worlds determined by whether it rains (R) and whether her credence in rain is high (H) or low (L), as shown in table 1.

	RH	RL	$\bar{R}H$	$\bar{R}L$
H	.3	.3	.2	.2
L	.2	.2	.3	.3

Table 1. Jane's credences depending on whether she's in an H or L world.

Suppose at every world, Jane is offered two bets. The first bet pays out \$1 if RL and costs 25¢. The second pays out \$1 if $\bar{R}H$ and also costs 25¢. Jane will only take the first bet when her credence is high and will only take the second bet when her credence is low, which will inevitably result in a loss, as shown in table 2.

The Dutch Book against Jane is a bit non-standard. Normally, a synchronic Dutch Book consists of a fixed set of bets such that any buyer of that book will lose money in any world. In other words, the bets bought remain invariant across worlds

[20] See Greaves and Wallace (2006).

[21] Some have taken the opacity of an agent's credences to motivate the idea that these credences may be imprecise. The advantages of self-confidence plausibly apply similarly to imprecise credence, but as imprecise decision theory is unsettled it's hard to prove a result analogous to Good's theorem.

	RH	RL	$\bar{R}H$	$\bar{R}L$
Bet 1	−25¢	declined	−25¢	declined
Bet 2	declined	−25¢	declined	−25¢
Total	−25¢	−25¢	−25¢	−25¢

TABLE 2. Jane's payouts.

and are guaranteed to lose. The book against Jane is different. Jane is offered the exact same options in every world, and those options do not combine in a guaranteed loss in and of themselves. Indeed, the set of bets Jane would accept in any particular world will never result in a guaranteed loss since Jane is a coherent EU-maximizer.[22] Instead, Jane is guaranteed a loss because which bets look fair or better to her vary from world to world, and this variance in her decision-making guarantees a loss.

More generally, we say that a bet is a fixed option if: (1) the bet costs the same at each world, and (2) the agent has the option to buy the bet at each world. An agent faces a **fixed-option Dutch Book** if there is some set of fixed options that she is offered which would result in a guaranteed loss in every world given her decisions as to whether to buy or turn down those bets. Thus, the book against Jane is a fixed option Dutch Book.[23]

As it turns out, any unconditional EU-maximizer is vulnerable to a fixed-option Dutch Book unless she puts a lot of stock in her first-order views. Let X be some random variable, such as the number of inches of rain in Seattle in the next ten years. Let $E(X) \geq x$ mean "the agent's expectation of X is at least x". Note that $E(X)$ is itself a random variable, since the agent might not know what her own expectation is if her credences are anti-luminous. It follows from BLINDED (thm. BLINDED) that the agent is vulnerable to a fixed-option Dutch Book if at some world ω, and for some random variable X and number x, $E_\omega(X \mid E(X) \geq x) < x$, where $E_\omega(X)$ is the expected value she assigns to X at ω. Such agents, we say, *do not trust themselves.*

In the example above, Jane does not trust herself. To see why, let c_L represent her credence function in worlds where her credence in rain is only .4, and let c refer to her credence function, whatever it is. Then $c_L(R \mid c(R) \geq .6) = .4$. That is, she thinks (in the Low worlds) "Given my credence in rain is high, I'm still only forty percent sure it will rain."

[22] This result was proved independently by Lehman (1955) and Kemeny (1955).

[23] Mahtani (2015) also discusses Dutch Books involving uncertainty about one's own credences. See Das (2020) and Dorst et al. (2021) for more on fixed option Dutch Books.

Now, one may think that it is good to trust yourself, and that self-trust is a rational requirement. However, when you are unsure what your credences are, you may not know enough about yourself to trust what you think. Your own credences may be as opaque to you as those of a stranger, and you may reasonably be suspicious of your own opinions. If you're a standard EU-maximizer, that leaves you vulnerable to Dutch Books, while self-confidence protects you from automatic exploitability.

So, generally, an EU-maximizing agent following a decision theory T should prefer to follow T's self-confident analogue. What, then, is not to like about self-confidence? Most obviously, self-confident decision functions seem completely divorced from the agent's point of view, and (traditionally) decision theory is all about the agent's point of view. Frank Jackson puts this sentiment well:

> [T]he fact that a course of action would have the best results is not in itself a guide to action, for a guide to action must in some appropriate sense be present to the agent's mind. We need, if you like, a story from the inside of an agent ... and having the best consequences is a story from the outside.[24]

A great virtue of decision theory is that it offers guidance, but agents performing strange feats of self-confidence will not understand the force that moves them. Self-confident agents may be completely unable to explain their actions. If given the opportunity to get a nickel, a self-confident agent will bet her life that she is in the exact decision-situation she is in. She may—and likely will—expect to die, and curse the inexplicable madness which led her to take the apparently insane bet.

A great virtue of decision theory is that it offers guidance, but that virtue is already compromised if we allow agents to be uncertain of their decision-situations. Any agent who is uncertain about her credences, utilities, and available acts is—by definition—less than fully conscious of the factors underpinning her decision. Standard decision theories are no better-suited for guidance than self-confident decision theories are. Consider an agent who has high credence that it's going to rain tomorrow, but who is wildly uncertain about what her credence about the rain is. Suppose this agent is offered a bet at even odds that it's going to rain tomorrow. This is a proposition that the agent thinks is true, so, standardly, the agent should take the bet. But the agent is in no position to be guided by that verdict. A standard decision theory makes its prescription regarding the bet about the rain on the basis of the agent's credence about the rain—but the agent is wildly

[24] Jackson (1991).

uncertain about what that credence is. Without luminosity, no decision theory can perfectly satisfy Jackson's ideal of guidance. There is no appropriate sense in which the prescriptions of a decision theory can always be present to an agent's mind.

Moreover, there is virtually no hope for traditional decision theory to retain guidance value when the agent is uncertain about her available acts. In traditional decision theory, acts are simply functions from states of the world to outcomes. Available acts are acts you could actually perform. Traditional decision theory ranks all acts, but it tells you to perform the *available* one with highest expected utility.

Suppose John sees $100 lying on the sidewalk beside him. John likes money, so he thinks he should bend down to pick it up. As he reaches for the bill, his back gives out, he fails to get the $100, and he spends three days in the hospital in traction.

John thought getting the $100 was an available act, but it actually wasn't. It wasn't something he could actually do at the point of decision. Of course, one might object that the act space shouldn't be conceived of as including things like "successfully bend down and pick something up". But no matter what 'normal' actions go in the act space, an agent might be uncertain as to what's available. You may be unsure whether you can successfully press a button in front of you, travel to Damascus, or tell a detective you want to cooperate.

It's unclear how we would even go about accommodating uncertainty about available actions in a principled way. Decision theory has to tell you what to do, so its verdict must be based on what you can do, not on your credences about what you can do. John can't roll a die that results in him bending down successfully 80% of the time, or reach only four-fifths of the way to the ground to grab $80 instead of $100. Unlike with credences, acts can't be hedged. Either you can do them or you can't.

A self-confident decision theory does not compromise guidance value; it merely takes advantage of a circumstance in which guidance value was already compromised. Any agent who is less than certain of her decision-situation must be something of a somnambulist, moving without understanding. Why not be a somnambulist who takes winning bets?

One might wonder if self-confident decision theory goes far enough. Why should agents be self-confident, and act as though they were sure of their actual decision-situations? Why not have agents be height-confident, and act as though they were sure of their actual heights? Or actuality-confident, and act as though they were sure of absolutely everything? We answer that self-confidence is meaningfully different than height-confidence or actuality-confidence. Self-confidence is formulable given only the standard apparatus of decision theory, whereas height-confidence and actuality-confidence require more apparatus. But there might well be circumstances in which more than the standard apparatus should be allowed. One might be able to build an agent whose actions are sensitive to her height, and in such a case it might be a good idea to make that agent height-confident. It would, of course, be rather harder to build an actuality-confident agent. But we don't balk at height-confidence or actuality-confidence. One needs to specify the domain of decision theory—to specify the tools that decision theorists can use when coming up with their theories—and only then can one determine which decision theory is appropriate. The standard formalism of decision theory does not allow for height-confidence or actuality-confidence, but it does allow for self-confidence.

6. Foundational Concerns

We've said that self-confident decision theories have a straightforward appeal: they better conduce to advantageous action than standard decision theories do. We've explained how to formulate self-confident decision theories using the credences, utilities, and available acts of standard decision theory. That's all well and good. Nonetheless, one might reasonably worry that it doesn't make sense for credences, utilities, and available acts to operate in the way that self-confident decision theory requires. If decision-situations of the sort we've discussed don't make philosophical sense, then self-confident decision theories are pointless.

It's well worth kicking the tires, and seeing whether the mathematical structures we've articulated have appropriate philosophical interpretations. Broadly speaking, there are two philosophical interpretations of credences and utilities: according to *non-constructivist realism*, credences and utilities are formal characterizations of agents' beliefs and desires, and as such have an existence independent of their preferences. According to *constructivism*, credences and utilities are representational devices defined by other mental states.[25] Each interpretation has a respectable history, and each interpretation poses a potential (though, we think, answerable) problem for self-confident decision theories. While we are inclined

[25] This terminology follows Buchak (2013)

toward non-constructivist realism and maintain it is at least a good model of certain artificial agents, we will not take a firm stand on the correct foundational approach here.[26]

6.1. **Non-Constructivist Realism.**

Suppose that credences and utilities characterize an agent's beliefs and desires. One might reasonably worry that agents' beliefs and desires are so transparent to them that the situations we have described—in which agents are uncertain about their credences and utilities—cannot occur.

This worry is especially forceful if credences and utilities correspond to an agent's conscious judgments, on the grounds that such judgments are luminous to agents. If your credence that p is the number you come up with when you reflectively decide how confident you are in a proposition, then identifying the number one came up with might well be unproblematic.[27]

Such thinking has precedent. A related line of reasoning is advanced in Berker (2008) to argue that a "constitutive connection" between believing that one is cold and being cold substantiates the luminosity of being cold. But Berker's principle only applies to one who has "done everything they can to decide whether one feels cold", and so doesn't apply when, say, an agent makes a snap judgment.[28] The sort of constitutive connection that could trivialize self-confident decision theory would have to be much stronger, something like

> One is certain that one's credence in p is x iff one's credence in p is x.

But a constitutive connection of this strength is implausible. It's contentious enough whether introspective certainty is guaranteed under ideal conditions; it is not guaranteed no matter what.

Moreover, even if agents' introspective abilities give them immaculate access to their credences and utilities, it would still be dubious to think that they have immaculate access to their available acts. Although it's easy enough to stipulate what acts are available in toy decision problems ('take umbrella', 'leave umbrella'),

[26] Reinforcement learning is a major paradigm in contemporary machine learning, and it operates according to a non-constructivist form. In reinforcement learning, an artificial agent learns to maximize an expected reward (explicitly represented as a floating point number) over time, and such an agent may also explicitly represent various probability functions such as its probability of taking a particular action in a particular state, or of seeing a state given a previous action and previous state Sutton and Barto (2018).

[27] Analogous considerations apply to utilities, though they are harder to state due to utilities being defined on an interval scale. We therefore focus on credences.

[28] See Spencer and Wells (2019) for examples of agents who are pressed for time and arguments about the decision-theoretic significance of that time pressure.

it's far from clear what the appropriate definition of "available act" is.[29] And it's even less clear what falls under such a definition in every possible circumstance.[30]

We think the case for anti-luminosity is especially strong regarding actions. Suppose that an agent has utilities that are linear in money, and so their credence in a proposition can be read straightforwardly off of the odds at which they are indifferent to a bet on that proposition. Must an agent be certain which bets they'll take and which they won't? We certainly don't think so. Similarly, an agent may have a thoroughly imperfect understanding of their own preferences in certain difficult situations. Some may think that they want to lose weight more than they want a tasty snack, and be surprised when they find themselves eating yet another cookie.[31] Some may think that they want to remain faithful to their spouse more than they want a passionate dalliance, and be surprised when they find themselves having an affair. Indeed, a standard reason why behavioral economists look to "revealed preferences" rather than self-reports is that people's descriptions of what they would do in a circumstance often vary from what they actually wind up doing in that circumstance.

6.2. **Constructivism.** Suppose that credences and utilities are representational devices defined by other mental states (such as preferences).[32] So long as these states conform to the appropriate formal requirements, one automatically counts as having corresponding credences and utilities. There are various extant theories for how credences and utilities should be construed, and various representation theorems have been proven.[33] According to all the extant theories, agents are represented as some sort of expected utility maximizer. But in that case, an agent who will bet on some proposition at any odds—no matter how unfavorable—must be certain that the proposition is true. However, we've said that an agent should

[29] See Jeffrey (1965), Lewis (1981), and Joyce (1999) for three prominent views.

[30] See Schwarz (2021) for an extended discussion of uncertainty regarding available acts.

[31] See Callard (2018).

[32] It's natural to wonder what we take preferences to be. We do not want to take a strong stand here, as the literature on such a question is vast, and self-confident decision theory is compatible with many different conceptions. When Jane refuses a bet about what her credences are at any odds despite her uncertainty, one may assume we endorse an account such as Sen's (1973), which takes her to act *against* her actual preference for taking the bet. However, given the framework of the representation theorem in the appendix, it is more natural for us to say she ultimately does not prefer to take the bet despite her credences. This accords with a view like Hausman's (2011), who takes preferences to be "total subjective comparative evaluations" (p. 4). In this instance, Jane's actual total subjective evaluation (if she's rational) would be not to take the bet, even if she is not aware that's her total evaluation.

[33] For example, Savage (1954), Jeffrey (1965), and Joyce (1999).

always bet that their decision-situation is as it actually is even when their corresponding credence is low. So if credences and utilities function in the standard way, then agents of the sort we have been describing cannot exist.

We have two rejoinders: (1) We maintain self-confident decision theories make representational sense regarding credences and utilities. (2) We maintain that self-confident decision theories make non-representational sense regarding available acts.

Regarding (1), we suggest that it is tendentious to presuppose that the appropriate representational schema is one of the standard, expected utility-maximizing ones. We are proposing novel decision theories after all; we can't very well presuppose that extant decision theories have the last word. While standard views do say that agents who are represented with middling credences towards a proposition would not bet on it at extremely unfavorable odds, the project of self-confident decision theory calls standard views into question.

The committed constructivist might have reservations about self-confident decision theories, on the grounds that they lack representation theorems. Representation theorems specify how credences and utilities can be characterized in other terms. Without a representation theorem, one might worry that a decision theory can't actually be appropriately cashed out.

To (partly) allay this concern, we show how to derive a representation theorem for self-confident EDT from the representation theorem for EDT in Joyce (1999). This approach could be extended to self-confident CDT by making adjustments analogous to those Joyce uses for his extension to CDT. The relevant details are rather technical, so we relegate them to an appendix.

Keep in mind that we're not only quibbling about the appropriate representation for particular choices. (We're doing that, but not only that.) Self-confident decision theory imposes substantive constraints on what courses of action are rationalizable. Standard decision theories would allow an agent to bet that their credences and utilities are a certain way at favorable odds and yet not take such a bet at unfavorable odds. Self-confident decision theories do not allow that pattern of choice: such bets must either be accepted at any odds or rejected at any odds, depending on whether the bets will win or lose.

Regarding (2), even if an agent's credences and utilities are representational devices, an agent's available acts are not. Suppose an agent is choosing between A, B, and C. This agent is, however, highly confident (mistakenly) that he can only choose between A and B. The thing is, the agent thinks that in circumstances in which he can choose between only A and B, it's better to go with A, but in circumstances in which he can choose between A, B, and C, it's better to go with B.[34] In these circumstances, choosing A is liable to maximize expected utility. But conditional on the fact that the agent can choose C, B maximizes expected utility. Even given constructivism, it's a fact that the agent can choose C, and one of which the agent is unaware. Even given constructivism, a self-confident decision theory that capitalizes on this information makes sense.

7. Reasoning by Cases

Because of the lack of luminosity no decision theory can always be followed with perfect reliability. One might nonetheless think that it's a problem that self-confident decision theories are strictly harder to follow than standard decision theories. But if so, one would be wrong. Self-confident decision theories are not strictly harder to follow than standard decision theories. In fact, there are famous cases in which self-confident decision theories are easier to follow than standard decision theories.

There are cases in which self-confident causal decision theory is easier to follow than standard causal decision theory. Consider the following––

> **The Psychopath Button:** Paul is debating whether to press the "kill all psychopaths" button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world with psychopaths to dying. Should Paul press the button?[35]

Let us suppose that Paul is initially quite confident that he is not a psychopath. Moreover, let us suppose that although Paul assigns higher utility to killing all psychopaths than allowing them to live (at least so long as he is not a psychopath himself) he's far from certain that he does. Paul also has substantial credence that he assigns higher utility to letting the psychopaths live rather than killing them, even if he is not a psychopath himself. For convenience, let's assume that Paul's credences are luminous; his only uncertainty about his decision situation is about the utility he assigns to killing psychopaths.

What does causal decision theory say Paul should do? Holding fixed Paul's beliefs about the causal structure of the world (in this case, whether or not he is a psychopath), the expected utility of pushing the button exceeds the expected utility of not pushing the button. Causal decision theory would thus have Paul

[34] Note that this does not violate the independence of irrelevant alternatives, as we do not presuppose act / state independence. For more on this issue, see Broome (1991), Ch. 5.

[35] Egan (2007).

push the button. But if Paul is a committed causal decision theorist it's hard for him to tell whether or not he should push the button. After all, Paul isn't sure what his utilities are, and so he can't be sure which of his actions maximizes causal expected utility.

What does self-confident causal decision theory say Paul should do? Holding fixed Paul's beliefs about the causal structure of the world and taking Paul's decision-theoretic state for granted, the expected utility of not pushing the button exceeds the expected utility of pushing the button. Moreover, even in his state of uncertainty about his utilities Paul can figure that fact out. Although Paul cannot straightforwardly calculate the self-confident expected utilities of his available acts, he doesn't have to. Paul can reason by cases instead. There are two possibilities: either (1) Paul assigns higher utility to killing all psychopaths than to letting them live, or (2) he assigns higher utility to letting psychopaths live than to killing them. Supposing (1) is true, Paul knows that the self-confident causal expected utility of pushing the button is low, because he'll know that conditional on those utilities his credence that he is a psychopath is high. Supposing (2) is true, Paul knows that the self-confident causal expected utility of pushing the button is low, because he'll know that conditional on those utilities he'd rather that the psychopaths live. If Paul is a self-confident causal decision theorist he can deduce that he ought not to push the button.

There are cases in which self-confident evidential decision theory is easier to follow than standard evidential decision theory. Consider the following––

> **The Medical Newcomb Problem:** Susan is debating whether to smoke. She believes that smoking is strongly correlated with lung cancer, but only because there is a common cause—a condition that tends to cause both smoking and cancer. Once we fix the presence or absence of this condition, there is no additional correlation between smoking and cancer. Susan prefers smoking without cancer to not smoking without cancer, and she prefers smoking with cancer to not smoking with cancer. Should Susan smoke?[36]

Let us suppose that Susan is offered a Camel cigarette. She knows that she prefers smoking any kind of cigarette to not smoking, but is uncertain whether she assigns higher utility to smoking Camel cigarettes or to smoking Marlboro cigarettes. If Susan smokes the Camel cigarette she is offered that will be evidence that she prefers Camels to Marlboros, and if she refuses the Camel cigarette that will be

[36] Egan (2007). But Egan takes the case from Gibbard and Harper (1978).

evidence that she prefers Marlboros to Camels. Moreover, Susan is highly confident that there's a common cause between the cancer-causing condition and one of the cigarette preferences—she's convinced that either people who prefer Camels or people who prefer Marlboros are especially likely to have this cancer-causing condition. But Susan isn't sure which cigarette preference is—in her own judgment—positively correlated with the cancer-causing condition. In fact, Susan prefers Camels to Marlboros and thinks that people who prefer Camels are at increased risk of having the cancer-causing condition, but she is uncertain both about her utilities and her credences.

What does evidential decision theory say Susan should do? Given Susan's credences and utilities, the expected utility of not smoking exceeds the expected utility of smoking. Evidential decision theory would thus have Susan not smoke. But if Susan is a committed evidential decision theorist it's hard for her to tell whether or not to smoke. After all, Susan isn't sure what her credences are, and so she can't be sure which of her actions maximizes evidential expected utility.

What does self-confident evidential decision theory say Susan should do? Taking Susan's decision-theoretic state for granted, the expected utility of smoking exceeds the expected utility of not smoking. Moreover, even in her state of uncertainty about her credences Susan can figure that fact out. Although Susan cannot straightforwardly calculate the self-confident expected utilities of her available acts, she doesn't have to. Susan can reason by cases instead. There are two possibilities: either (1) Susan's cigarette brand preference strongly correlates with the underlying, cancer-causing condition, or (2) Susan's cigarette brand preference doesn't correlate with the underlying, cancer-causing condition. Given either (1) or (2) Susan's credences conditional on her decision-theoretic state will show no correlation between her cigarette brand preferences and whether or not she smokes; conditional on her actual decision-theoretic state her credences about her preferences will all be either 1 or 0, and any proposition with probability 1 or 0 is probabilistically independent of any other proposition. The self-confident evidential expected utility of smoking is thus guaranteed to exceed that of not smoking, given the mere fact that—other things being equal—Susan would rather smoke than not. If Susan is a self-confident evidential decision theorist she can deduce that she ought to smoke.

The statuses of *The Psychopath Button* and *The Medical Newcomb Problem* are, of course, controversial. They are also underdescribed. There are myriad ways to precisify what's at stake in them; the suppositions we have used to flesh out those cases are not the only ones possible. Really, there are many *Psychopath Buttons* and many *Medical Newcomb Problems*, and none of them are fully agreed-upon.[37] But it is clear that in the above precisifications, the dictates of self-confident decision theories are easier to follow than those of traditional decision theories.[38]

8. Conclusion

Causal decision theory and evidential decision theory each violate Anti-Futility, prescribing actions which are ill-suited for the decision-situations for which they are prescribed. These violations all stem from circumstances which exploit an agent's ignorance about his decision-situation. But self-confident decision theorists cannot be so exploited, even under conditions of uncertainty. Self-confident causal decision theory and self-confident evidential decision theory are not subject to counterexample as their standard analogues are. The conflict between causal decision theory and evidential decision theory should thus be superseded by a conflict between self-confident causal decision theory and self-confident evidential decision theory.

Appendix: Representation Theorem

We show how to extend a representation theorem for standard Evidential Decision Theory to self-confident Evidential Decision Theory. Similar techniques can extend this theorem to self-confident CDT, but we leave out those details here.

Our approach is derived from Joyce (1999) who takes both *comparative* probability along with preference as fundamental. We write $X \succeq Y$ to mean the agent regards X as at least as likely as Y, and $X \trianglerighteq Y$ to mean the agent regards X as more desirable/preferable to Y. Predictably, we write $X \succ Y$ ($X \vartriangleright Y$) to mean the agent regards X as strictly more likely (strictly preferable to) Y, and $X \approx Y$ ($X \bowtie Y$) to mean the agent regards X and Y as equally likely (is indifferent between X and Y).

Joyce takes $\succeq$ as a quasi-autonomous attitude for reasons that (of course) have nothing to do with self-confidence. Joyce notes that the standard Jeffrey-Bolker axioms (that appeal only to $\trianglerighteq$) famously fail to determine a unique probability function and unique up-to-affine transformation utility function that represent the

[37] Arif Ahmed has presented several ingenious arguments that it is not always irrational to push the "kill all psychopaths button". But if pushing the "kill all psychopaths" button may or may not be rational depending on how the case is precisified, then causal decision theory (which always prescribes pushing the button) is still refuted. Thus Ahmed's arguments are, at best, an incomplete defense of causal decision theory. For more see Ahmed (2012) and Ahmed (ms).

[38] We're thinking of following a decision theory as merely being a matter of doing what it says to do, regardless of why. We take it that the ease of discerning what self-confident decision theory says to do in these cases will make it easier to follow self-confident decision theory in this sense.

agent. Indeed, these axioms even allow agents to have preferences representable only by probabilistically incoherent likelihood rankings (Joyce, 1999, p. 136). So, although we admit that invoking a separate notion of comparative confidence may offend austere constructivists, such an approach is independently motivated.

Here is the basic idea underlying our theorem: At each state in the state space, the agent has a preference and comparative confidence ranking. The comparative confidence ranking allows the agent to be uncertain about her own levels of confidence. We require her preference ranking, on the other hand, to treat any proposition as null if the elements of that proposition all come with a different confidence or preference ranking.

Joyce's EDT Representation Theorem. With Joyce, we will define a decision-situation as a tuple $D = \langle \Omega, \mathcal{F}, O, S, \mathcal{A} \rangle$. O is the set of outcomes, S is the set of states, and $\mathcal{A}$ is the set of acts. Ω is the set of fully determinate possible worlds with ω. $\mathcal{F}$ is an algebra over Ω. Each element of O, S, and $\mathcal{A}$ is also in $\mathcal{F}$.

Joyce's EDT axioms over preferences are effectively the same as the Jeffrey/Bolker axioms. We list them, with slight modification, here.

EDT1: For some $G \in \mathcal{F}$, $G \rhd G \vee \neg G \rhd \neg G$.

EDT2: The decision-maker's preferences totally order the propositions in $\mathcal{F}$.

EDT3: If X and Y are mutually incompatible propositions and $X \unrhd Y$, then $X \unrhd X \vee Y \unrhd Y$.

EDT4: If $X \unrhd X'$ and if Y and Z are both incompatible with X and with X', then the following pattern of preferences never occurs unless each of the weak preferences is an indifference:

$$Y \rhd X' \vee Y \unrhd X \vee Y \rhd X \unrhd X' \rhd X' \vee Z \unrhd X \vee Z \rhd Z.$$

We refer the interested reader to (Joyce, 1999, pp. 128ff) for in depth motivation for and discussion of these and later axioms.

Again following Joyce, we define the following definition of coherence between $\unrhd$ and $\succeq$.

Coherence: We say that $\unrhd$ *coheres* with $\succeq$ if for any $X, X', Y \in \mathcal{F}$ such that Y is incompatible both with X and with X', we have:

- $X \succeq X'$ holds when either $Y \rhd X' \vee Y \unrhd X \vee Y \rhd X \unrhd X'$ or $X' \unrhd X \rhd X' \vee Y \unrhd X \vee Y \rhd Y$, and $X \succ X'$ holds when at least one $\unrhd$ in these chains is replaced with $\rhd$.
- $X \approx X'$ holds when $Y \rhd X' \vee Y \bowtie X \vee Y \rhd X \bowtie X'$ or $X' \bowtie X \rhd X' \vee Y \bowtie X \vee Y \bowtie Y$.

We also define nullity:

Nullity: X is *null* relative to $\trianglerighteq$ if and only if $X \vee Y \bowtie Y$ for some $Y \in \mathcal{F}$ that is incompatible with X and for which either $X \triangleright Y$ or $Y \triangleright X$.

The final core preference axioms require treating null propositions consistently as, in effect, probability zero events:

EDT5: If X is null, then $X \vee Y \bowtie Y$ for all $Y \in \mathcal{F}$.

EDT6: $\bot$ is null, and if $X \in \mathcal{F}$ is null, then $X \wedge Y$ is null for all $Y \in \mathcal{F}$.

We can also add the following four axioms to ensure only countably additive probabilities will represent the decision-maker.

EDT7: If $\{X_1, X_2, \ldots\}$ is a countable and each X_i is null, then $X = \bigvee_{i=1}^{\infty} X_i$ is null.

EDT8: Any collection of pairwise incompatible non-null propositions is countable.

EDT9: Let $\{X_1, X_2, \ldots\}$ be a countable set of pairwise incompatible propositions and $X = \bigvee_{i=1}^{\infty} X_i$.

- If $Y \geq \bigvee_{i=1}^{n} X_i$ for all n, then $Y \geq X$
- If $\bigvee_{i=1}^{n} X_i \geq Y$ for all n, then $X \geq Y$.

EDT10: If $X \in \mathcal{F}$ is non-null, then there exist incompatible non-null propositions $X_1, X_2 \in \mathcal{F}$ such that $X = X_1 \vee X_2$.

Note that this last axiom ensures that $\mathcal{F}$ is non-atomic.

These axioms are enough for the standard Jeffrey/Bolker representation theorem. However, Joyce (1999, p. 138) also adds a list of comparative probability axioms strong enough on their own to ensure a unique countably additive probabilistic representation. We won't reproduce those here but instead will just refer to them as the CP-axioms.

More explicitly, we have:

Theorem 1 (Villegas)**.** If $\geq$ satisfies the CP axioms, there exists a unique countably additive probability function P defined on $(\Omega, \mathcal{F})$ such that for all $X, Y \in \mathcal{F}$, $P(X) \geq P(Y)$ iff $X \geq Y$.

We then get Joyce's representation theorem:

Theorem 2 (Joyce)**.** Let $\langle \Omega, \mathcal{F}, O, S, \mathcal{A} \rangle$ be a decision frame with $\mathcal{F}$ a σ-algebra over Ω. If $\geq$ and $\trianglerighteq$ are defined over $\mathcal{F}$, $\geq$ obeys the CP-axioms, $\trianglerighteq$ obeys the EDT axioms, and $\geq$ coheres with $\trianglerighteq$, then there exists a unique pair (P, u) consisting of a countably additive probability P on $(\Omega, \mathcal{F})$ and a real-valued utility u defined on O such that:

(1) P represents $\succeq$,
(2) The conditional expected utility function $V(X) = \sum_{o \in O} P(o \mid X) u(o)$ both represents $\trianglerighteq$ and obeys the scaling convention that $V(G) = 1$ and $V(\top) = 0$, where G is some pre-designated proposition such that $G \succ \neg G$.

We can now turn to extending this theorem in a natural way to allow for uncertainty about one's own decision-situation.

Self-Confident EDT Representation. With Joyce, we assume a space of possibilities Ω and algebra over the possibilities $\mathcal{F}$ and states S. For any $\omega \in \Omega$, there is a set of outcomes O_ω. We let $\mathcal{O} := \{O_\omega \mid \omega \in \Omega\}$. We also require that every element of $\mathcal{A}$ and S is in $\mathcal{F}$, as is each element of O_ω for every ω. So, a decision *frame* is a tuple $D = \langle \Omega, \mathcal{F}, \mathcal{O}, S, \mathcal{A} \rangle$.

We also need to equip each $\omega \in \Omega$ with its own preference ranking $\trianglerighteq_\omega$ and comparative confidence ranking $\succeq_\omega$. We can then define the equivalence class $d(\omega) := \{\omega' \in \Omega \mid \succeq_\omega = \succeq_{\omega'}, \trianglerighteq_\omega = \trianglerighteq_{\omega'}\}$.

Our new comparative probability axioms are the same as Joyce's. That is, we require that for each $\omega \in \Omega$, $\succeq_\omega$ obeys the usual comparative confidence axioms. We also require that

CP-T: for every $\omega \in \Omega$, $d(\omega) \succ_\omega \bot$

This axiom ensures that the agent is represented at each world ω by a probability function P_ω such that $P_\omega(\cdot \mid d(\omega))$ is defined.

Given the other CP axioms, Villegas's theorem then ensures that the agent is representable at each ω by a unique countably additive probability function P_ω defined over $\mathcal{F}$.

The preference axioms also require only modest changes. In most of the standard axioms, we can simply replace $\triangleright$ and $\trianglerighteq$ with $\triangleright_\omega$ and $\trianglerighteq_\omega$ (along with the obvious changes).

We need to change EDT1 to ensure that a non-trivial proposition exists within $d(\omega)$ for each $\omega \in \Omega$.

SCEDT-1: For every $\omega \in \Omega$, there exists $G \in \mathcal{F}$ such that $G \subset d(\omega)$ and $G \triangleright_\omega G \vee \neg G \triangleright_\omega \neg G$.

We also require that the agent treat all propositions incompatible with $d(\omega)$ as null at $d(\omega)$. That is, we add:

SCEDT-Nullity: For any $\omega \in \Omega$ and $X \in \mathcal{F}$, if $X \cap d(\omega) = \emptyset$, then X is null according to $\trianglerighteq_\omega$.

So, in effect, the comparative confidence axioms allow higher-order uncertainty since they permit the agent to treat states with different decision situations as possible, but the preference axioms require her to treat those states as null. (Note that this additional nullity axiom is kosher because it requires her to treat certain states as null based on factors internal to her decision situation, not based on outside empirical knowledge.)

The only missing piece is an analogue of the definition of coherence between $\succeq$ and $\trianglerighteq$. For self-confident EDT, $\succeq_\omega$ won't in general cohere with $\trianglerighteq_\omega$, since the former may treat propositions incompatible with $d(\omega)$ as more likely than the contradiction, but $\trianglerighteq_\omega$ treats them as null.

Instead, what we care about is the agent's opinions *conditional* on her decision situation. To do so, we first define the ranking $\succeq^d_\omega$ as the conditional comparative confidence ranking given $d(\omega)$. In other words, $X \succeq^d_\omega Y$ if and only if $X \wedge d(\omega) \succeq_\omega Y \wedge d(\omega)$. We can then define the notion of coherence relevant to self-confident EDT as follows:

> **Luminous Coherence:** Let $\omega \in \Omega$. We say $\succeq_\omega$ *luminously coheres* with $\trianglerighteq_\omega$ if and only if $\succeq^d_\omega$ coheres with $\trianglerighteq_\omega$.

We can then straightforwardly derive the representation theorem for Self-Confident EDT.

Theorem 3. Let $\langle \Omega, \mathcal{F}, \mathcal{O}, S, \mathcal{A} \rangle$ be a decision-frame with $\mathcal{F}$ a σ-algebra over Ω. If $\succeq_\omega$ and $\trianglerighteq_\omega$ are defined over $\mathcal{F}$, $\succeq_\omega$ obeys the CP-axioms along with CP-T, $\trianglerighteq_\omega$ obeys the EDT axioms along with SCEDT-1 and SCEDT-Nullity, and $\succeq_\omega$ luminously coheres with $\trianglerighteq_\omega$, then there exists a unique pair (P_ω, u_ω) consisting of a countably additive probability P_ω on $(\Omega, \mathcal{F})$ and a real-valued utility u_ω defined on O_ω such that:

(1) P_ω represents $\succeq_\omega$
(2) The conditional expected utility function $V(X) = \sum_{o \in O_\omega} P(o \mid X, d(\omega)) u_\omega(o)$ both represents $\trianglerighteq_\omega$ and obeys the scaling convention that $V(G) = 1$ and $V(\top) = 0$, where G is some pre-designated proposition in $d(\omega)$ such that $G \succ \neg G$.

The reason this theorem holds is straightforward, given Joyce's theorem. By the definition of luminous coherence and Joyce's theorem, we know that $P(\cdot \mid d(\omega))$ represents $\succeq^d_\omega$ and that $\succeq^d_\omega$ coheres with $\trianglerighteq_\omega$. Since $P(\cdot \mid d(\omega))$ represents our agent's self-confident credences at ω, she is then representable as a self-confident EDT-agent.

References

Ahmed, A. (2012). Push the button. *Philosophy of Science 79*(3), 386–395.

Ahmed, A. (ms). Smokers and psychos: Egan cases don't work.

Berker, S. (2008). Luminosity regained. *Philosophers' Imprint 8*, 1–22.

Bonjour, L. (1980). Externalist theories of empirical knowledge. *Midwest Studies in Philosophy 5*(1), 53–73.

Broome, J. (1991). *Weighing Goods: Equality, Uncertainty and Time.* Wiley-Blackwell.

Buchak, L. (2013). *Risk and Rationality.* Oxford University Press.

Callard, A. (2018). *Aspiration: The Agency of Becoming.* Oup Usa.

Carr, J. R. (2020). Imprecise evidence without imprecise credences. *Philosophical Studies 177*(9), 2735–2758.

Das, N. (2020). Externalism and exploitability. *Philosophy and Phenomenological Research.*

Dorst, K., B. A. Levinstein, B. Salow, B. E. Husic, and B. Fitelson (2021). Deference done better. *Philosophical Perspectives 35*(1), 99–150.

Egan, A. (2007). Some counterexamples to causal decision theory. *Philosophical Review 116*(1), 93–114.

Forster, E. (1985). *Aspects of the Novel.* Clark lectures. Harcourt Brace Jovanovich.

Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies 156*(1), 33–63.

Gibbard, A. and W. L. Harper (1978). Counterfactuals and two kinds of expected utility. In A. Hooker, J. J. Leach, and E. F. McClennen (Eds.), *Foundations and Applications of Decision Theory*, pp. 125–162. D. Reidel.

Good, I. (1967). On the principle of total evidence. *The British Journal for the Philosophy of Science 17*, 319–322.

Greaves, H. and D. Wallace (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind 115*(459), 607–632.

Hájek, A. (2016). Deliberation welcomes prediction. *Episteme 13*(4), 507–528.

Hausman, D. M. (2011). *Preference, value, choice, and welfare.* Cambridge University Press.

Jackson, F. (1991). Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics 101*(3), 461–482.

Jeffrey, R. C. (1965). *The Logic of Decision.* New York, NY, USA: University of Chicago Press.

Joyce, J. M. (1999). *The Foundations of Causal Decision Theory.* Cambridge University Press.

Joyce, J. M. (2002). Levi on causal decision theory and the possibility of predicting one's own actions. *Philosophical Studies 110*(1), 69–102.

Kagan, S. (2018). The paradox of methods. *Politics, Philosophy and Economics 17*(2), 148–168.

Kemeny, J. G. (1955). Fair bets and inductive probabilities. *Journal of Symbolic Logic 20*(3), 263–273.

Lasonen-Aarnio, M. (forthcoming). Perspectives and good dispositions. *Philosophy and Phenomenological Research.*

Lehman, R. S. (1955). On confirmation and rational betting. *Journal of Symbolic Logic 20*(3), 251–262.

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy 59*(1), 5–30.

Maher, P. (1990). Symptomatic acts and the value of evidence in causal decision theory. *Philosophy of Science 57*(3), 479–498.

Mahtani, A. (2015). Dutch books, coherence, and logical consistency. *Noûs 49*(3), 522–537.

Peacocke, C. (1998). *Being Known.* Oxford University Press.

Ramsey, F. P. (1926). Truth and probability. In H. E. Kyburg and H. E. Smokler (Eds.), *Studies in Subjective Probability.* Huntington, NY: Robert E. Krieger Publishing Co.

Savage, L. J. (1954). *The Foundations of Statistics.* Wiley Publications in Statistics.

Schoenfield, M. (2017). Conditionalization does not maximize expected accuracy. *Mind 126*(504), 1155–1187.

Schwarz, W. (2021). Objects of choice. *Mind.*

Sen, A. (1973). Behaviour and the concept of preference. *Economica 40*(159), 241–259.

Spencer, J. and I. Wells (2019). Why take both boxes? *Philosophy and Phenomenological Research 99*(1), 27–48.

Sutton, R. S. and A. G. Barto (2018). *Reinforcement learning: An introduction.* MIT press.

Williamson, T. (2000). *Knowledge and its Limits.* Oxford University Press.

Williamson, T. (2008). Why epistemology cannot be operationalized. In Q. Smith (Ed.), *Epistemology: New Essays.* Oxford University Press.